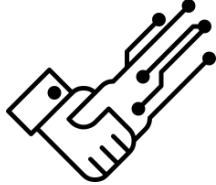
ИСП РАН



Обеспечение безопасности глубоких моделей классификации в условиях OOD и атак уклонения

Лукьянов Кирилл Сергеевич
м.н.с. центра искусственного интеллекта ИСП РАН

IV форум технологии доверенного искусственного интеллекта
27.05.2024 Москва

Введение

В более общем смысле можно определить носитель с заданным уровнем вероятности $\tau \in [0, 1)$ для дискретного случая и $\tau \in [0, \infty)$ для непрерывного случая :

$$support(\mathbb{P}) = \{x \in X | \mathbb{P}(X = x) > \tau\}$$

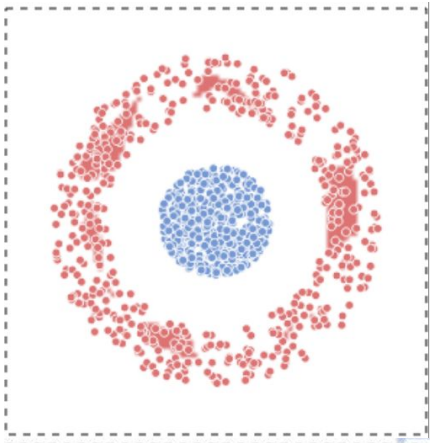
и

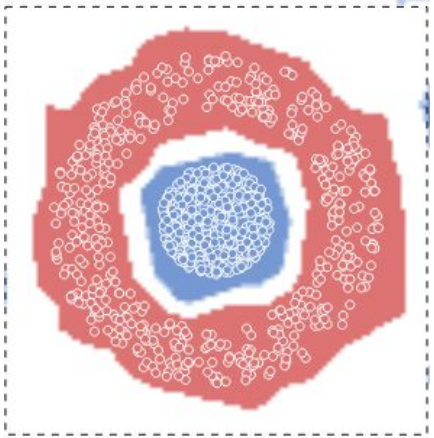
$$support(\mathbb{P}) = \overline{\{x \in X | f_X(x) > \tau\}},$$

В случае конечных наборов данных приближенный носитель распределения определяется следующим образом

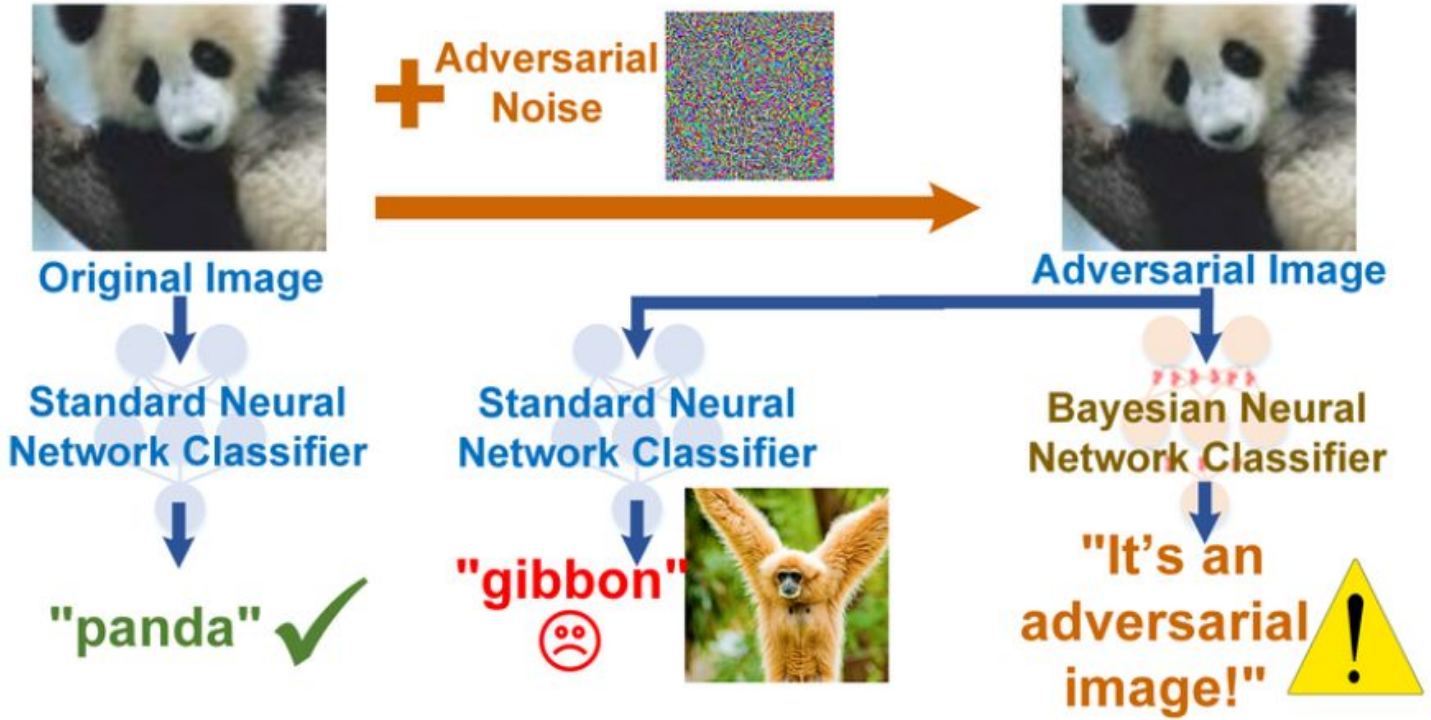
$$\hat{support}(\mathbb{P}) = \overline{\{x \in X | \hat{f}_X(x) > \tau\}},$$

где $\hat{f}_X(x)$ – оценка функции плотности вероятности.



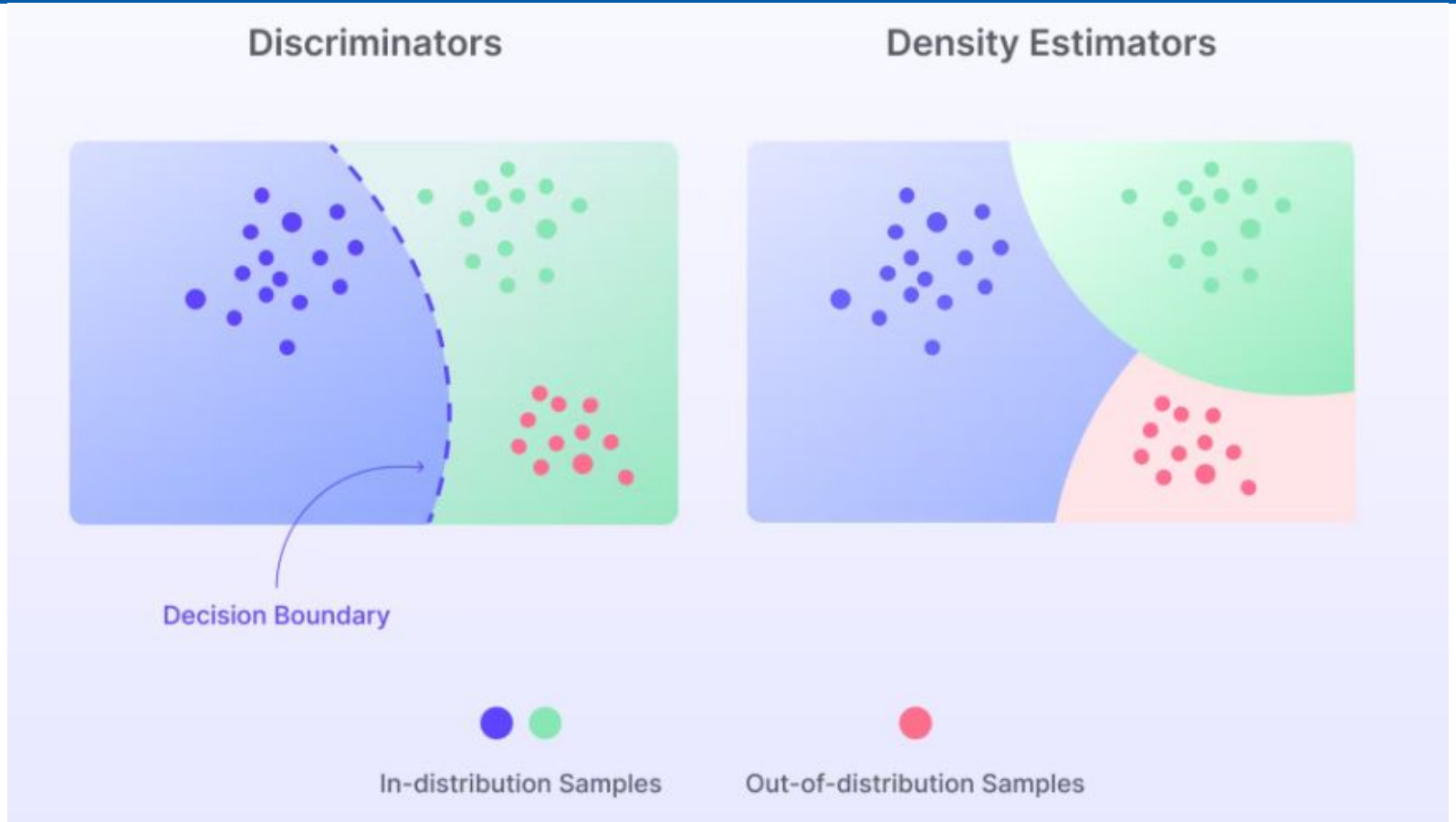


Защищенность



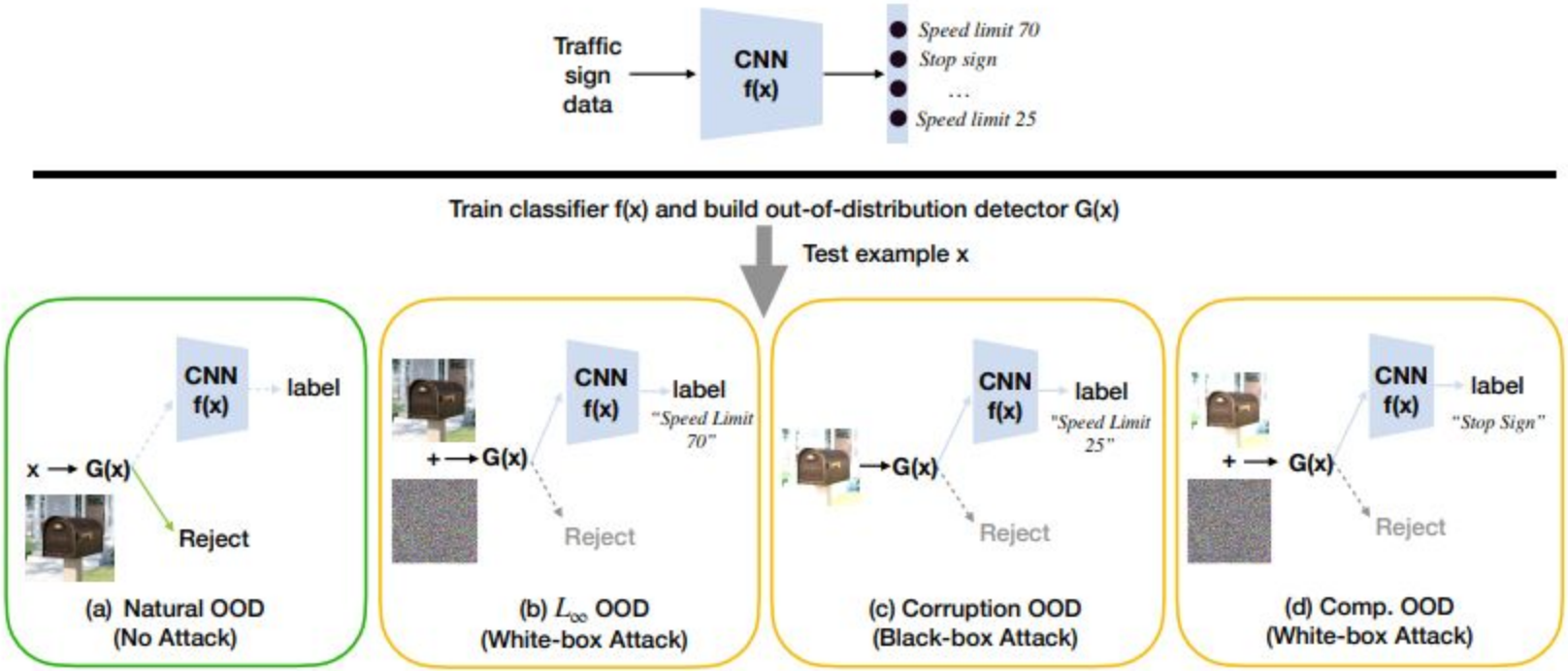
https://www.researchgate.net/figure/An-original-image-added-with-small-perturbations-that-are-imperceptible-to-human-eyes_fig1_339266924

Корректность



https://encord.com/blog/what-is-out-of-distribution-ood-detection/

Дополнительные задачи при обеспечении безопасности



https://hackmd.io/@tungsomot/Bk4Biy2Hi

Доверенный ИИ и связь с задачей выделения носителя

- Интерпретируемость
- Защищенность
- Корректность
- Приватность
- Справедливость
- Подотчетность
- Другие

Какие проблемы решаются при решении задачи выделения носителя:

1. Методы интерпретации (аспект интерпретации)
2. Методы повышения защищенности от атак уклонения (аспект защищенности)
3. Методы детекции примеров вне распределения (аспект корректности)
4. Методы детекции отсутствия дрейфа данных (аспект корректности)
5. Методы детекции примеров вне распределения подкрепленных атаками (аспект корректности + защищенности)

Выделение носителя в пространствах малой размерности

Методы выделения носителя

1. Геометрические методы
2. Непараметрические статистические методы
3. Параметрические модели
4. Современные методы аппроксимации функции плотности распределения (VAE, Байесовские модели)

Основные ограничения 1-3 групп:

1. Полная или практически полная невозможность применения в пространствах с размерностью выше 50
2. Невозможность работать со сложными структурными объектами (текст, графы и т.д.)
3. Высокая зависимость от корректной нормализации данных
4. Нет адаптации для защиты от атак

Основные ограничения 4 групп:

1. Ограниченная математической база при принятии решении, результат является больше эмпирическими
2. Ограниченная поддержка сложных структурных объектов
3. Высокая стоимость обработки данных во время использования подобных техник

Выделение носителя в пространствах малой размерности

- **Методика выделения доверенных областей.**
  1. Предобработка данных (метки ± 1).
  2. Нормализация (опционально).
  3. Построение компакта. Для каждого признака определяются минимальное и максимальное значения, границы компакта выбираются на 25% шире относительно размаха.
  4. Генерация равномерного шума в пределах компакта (метка 0).
  5. Обучение нейронной сети.
  6. Выбор минимальной регрессионной оценки для принятия решения (β).
  7. Построение дерева eXBTree (критерий остановки: либо в листе остались точки одного класса и шумовые точки, либо была достигнута максимальная глубина, соответствующая количеству нейронов в модели).
  8. Для каждой заполненной ячейки рассчитывается величина

$$R_{\text{leaf}} = \frac{N_{y=1} - N_{y=-1}}{N_{y=1} + N_{y=-1} + N_{\text{noise}}},$$

где $N_{y=1}$ – количество точек с меткой $+1$ в ячейке, а $N_{y=-1}$ и N_{noise} определяются аналогично для соответствующих меток.

Ячейки с $N_{y=1} + N_{y=-1} + N_{\text{noise}} < N_{\text{filled}}$ считаются *незаполненными*, где N_{filled} – дополнительный гиперпараметр, задающий минимальное количество точек в ячейке, при котором она считается заполненной.

Если ячейка незаполненная, т. е. не содержит достаточного числа точек для расчета R_{leaf}, но в ячейку попала точка, которую необходимо классифицировать, то для оценки этой ячейки (и классификации точки) необходимо подниматься по дереву послойно от листа к корню до тех пор, пока в ячейке текущего слоя не окажется столько точек, сколько необходимо для расчета R_{leaf}. (С геометрической точки зрения это эквивалентно объединению мелких ячеек в более крупную до тех пор, пока в этой ячейке не появится информация, на основании которой можно принять решение, является ячейка доверенной или нет.)

  9. Ячейки с $R_{\text{leaf}} > \beta$ считаются *доверенными* ячейками класса $+1$, ячейки с $R_{\text{leaf}} < -\beta$ – *доверенными* ячейками класса -1, остальные ячейки – *недоверенные*.

- **Расширение для решения практических проблем.**
  1. *Приложение для поиска "выбросов".* Если точка попала в незаполненную ячейку, которая в процессе "укрупнения" пополняется в основном точками "фона", то эту точку можно считать "выбросом".
  2. *Приложение для OOD.* Если для классификации приходит некоторый набор данных, значительная часть которого попадает в одни и те же незаполненные или заполненные в основном "фоном" ячейки, то можно считать, что имеет место "дрейф" распределения данных (что́ считать "значительной" частью, необходимо дополнительно определить, основываясь на практическом опыте, на предположениях об априорном распределении классов или специфике предметной области).
  3. *Детекция зон пересечения распределений классов.* Если

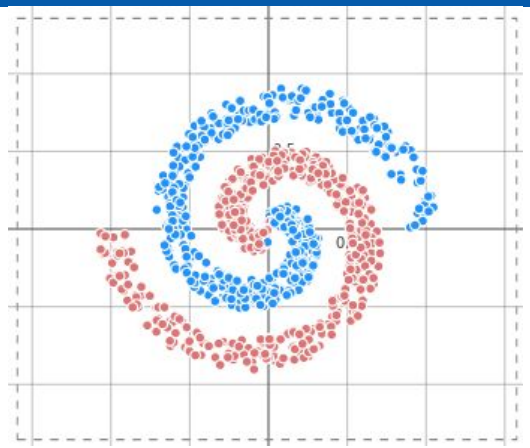
$$R_{\text{leaf}+} = \frac{N_{y=1}}{N_{y=1} + N_{y=-1} + N_{\text{noise}}} > \beta$$

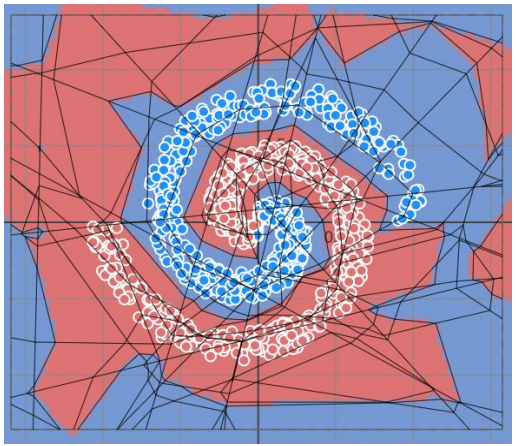
и

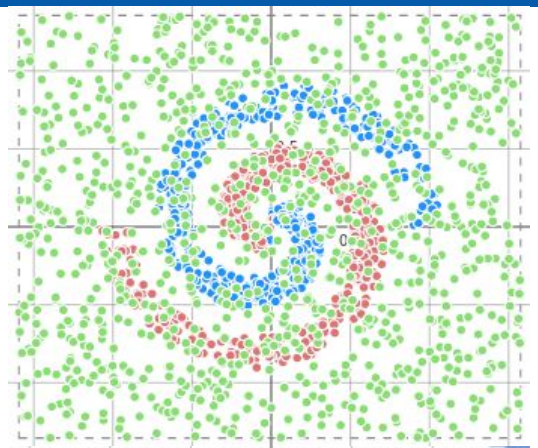
$$R_{\text{leaf}-} = \frac{-N_{y=-1}}{N_{y=1} + N_{y=-1} + N_{\text{noise}}} < -\beta,$$

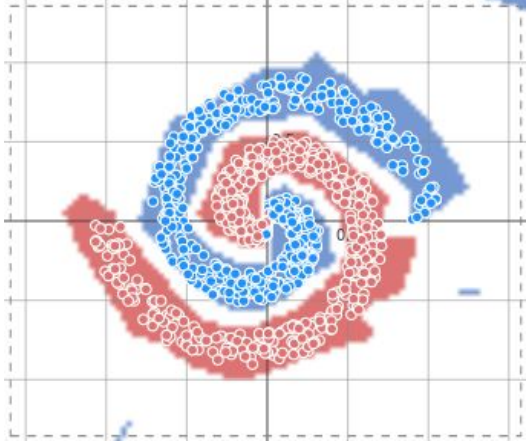
то эту ячейку можно определить как область пересечения распределений классов и при текущих обучающих данных разделить точки классов в этой ячейке невозможно.

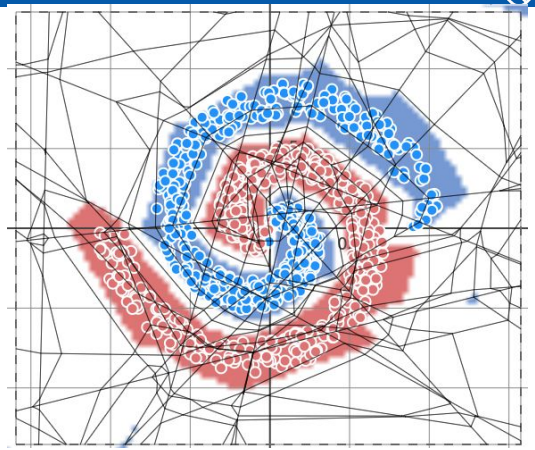
Выделение носителя в пространствах малой размерности











$$R_{leaf} = \frac{N_{y=1} - N_{y=-1}}{N_{y=1} + N_{y=-1} + N_{noise}}$$

где $N_{y=1}$ — количество точек с меткой +1 в ячейке, $N_{y=-1}$ и N_{noise} аналогично для соответствующих меток.

Выделение носителя в пространствах малой размерности

Таблица 10 — Качества детекции OOD и снижение эффективность атаки (в среднем по 20 запускам)

Тип обучения	ASR (↓)	Acc. OOD (↑)
Обычное обучение; 7 призн.	$98.2 \pm 1.3\%$	0%
С шумом; 7 призн.	$10.2 \pm 1.5\%$	$98.7 \pm 1.8\%$
Обычное обучение; 20 призн.	$99 \pm 0.5\%$	0%
С шумом; 20 призн.	$11.5 \pm 2.1\%$	$96.5 \pm 2.7\%$

Выделение носителя в пространствах высокой размерности

Вход алгоритма: $D = \{(X_i, Y_i)\}_{i=1}^{n}$ – обучающая выборка; E_0 и E_1 – число эпох начального обучения и число эпох обучения доверенного классификатора; параметр расширения $\rho > 0$ (напр., $\rho = 0{,}1$); m – число шумовых точек. **Выход алгоритма:** тройка $f : \mathbb{R}^d \to \mathbb{R}^J$, $g : \mathbb{R}^J \to \Delta_K$ и $p : \mathbb{R}^J \to \Delta_1$, определяющая классификатор $g \circ f$ и оценка вероятности p, что данный набор представлений не соответствует доверительной области.

Инициализация: задать параметрическую модель $\mathcal{M}$ классификаторов $h = g \circ f$, где $f : \mathbb{R}^d \to \mathbb{R}^J$ и $g : \mathbb{R}^J \to \Delta_K$, задать параметрическую модель для $p : \mathbb{R}^J \to [0,1]$.

Цикл от $t = 1$ **до** E_0**:** обучить f,g по D.

Вычислить представления $Z_i = (Z_{ij})_{j=1}^{J} = f(X_i)$, $i = 1, \ldots, n$.

Сгенерировать случайную выборку $\{U_i\}_{i=1}^{m}$ из равномерного распределения на $\Pi = \prod_{j=1}^{J} [m_j - \delta_j, M_j + \delta_j]$ для $m_j = \min_i Z_{ij}$, $M_j = \max_i Z_{ij}$ и $\delta_j = \rho(M_j - m_j)$.

Зафиксировать f и задать обучающую выборку $\bar{D} = \{(Z_i, 1)\}_{i=1}^{n} \cup \{(U_i, 0)\}_{i=1}^{m}$ в пространстве представлений, присвоив всем U_i метку 0 (шум), а Z_i метку 1 (начальные данные).

Цикл от $t = 1$ **до** E_1**:** Обучить по $\bar{D}$ функцию $p(z) \in [0,1]$, задающую условную вероятность класса 0 для всякой точки $z \in \mathbb{R}^J$, положив $p(z) = 1$ для $z \notin \Pi$ (в таких точках функция $p(z)$ принимает любое значение, при этом $p(z) = 1$ отвечает недоверенной зоне).

Объект из пространства высокой размерности

Экстрактор признаков

Переход, обычно соответствует размерности выхода

Классификатор модели

Выходная размерность соответствует количеству классов

Выделение носителя в пространствах высокой размерности

Таблица 11 — Оценка снижение эффективности атаки (в среднем по 20 запускам) на наборе данных CIFAR-10, модель ResNet-18. Все метрики представлены в процентах (%)

Используемый метод	Точность (↑)	ASR-PGD (↓)
Обычная модель	88 ± 1.2	99.8 ± 0.1
TRADES	73 ± 0.5	21.7 ± 1.0
IGR	**83.3** ± 1.5	83.2 ± 0.1
ODIN	42.2 ± 4	98 ± 0.2
DEN	**80.6** ± 3.0	**14.3** ± 3.5

Таблица 12 — Точность детекции OOD (в среднем по 20 запускам) на наборе данных CIFAR-10, модель ResNet-18. Все метрики представлены в процентах (%)

Используемый метод	Точность (↑)	OOD-noise (↑)	OOD-class (↑)	OOD-PGD (↑)
Обычная модель	88 ± 1.2	0 ± 0	0 ± 0	0 ± 0
TRADES	73 ± 0.5	0 ± 0	0 ± 0	0 ± 0
IGR	**83.3** ± 1.5	0 ± 0	0 ± 0	0 ± 0
ODIN	42.2 ± 4	97.6 ± 1.2	83.4 ± 2.6	2.4 ± 1.1
DEN	**80.6** ± 3.0	**100** ± 0	**89.8** ± 1.5	**83.4** ± 2.2

Таблица 13 — Точность детекции OOD и снижение эффективности атаки (в среднем по 20 запускам) на наборе данных Cora, модель GCN-2l. Все метрики представлены в процентах (%)

Используемый метод	Точность (↑)	ASR (↓)	OOD-noise (↑)	OOD-class (↑)	OOD-PGD (↑)
Обычная модель	87 ± 2	35 ± 4	0 ± 0	0 ± 0	0 ± 0
AdvTrain	**85** ± 6	**12** ± 5	0 ± 0	0 ± 0	0 ± 0
IGR	**84** ± 5	23 ± 4	0 ± 0	0 ± 0	0 ± 0
ODIN	55 ± 4	30 ± 3	90 ± 2	6 ± 7	10 ± 5
DEN	**79** ± 4	**13** ± 4	**100** ± 0	**71** ± 4	**65** ± 7

Выделение носителя в пространствах высокой размерности

Теорема

Пусть $X \sim \mathcal{N}(\mu, \Sigma)$, где $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ положительно определена. Пусть $f : \mathbb{R}^d \to \mathbb{R}^n$ – липшицева функция, $J_f(x) \in \mathbb{R}^{n \times d}$ – её матрица Якоби (существует почти всюду по теореме Радемахера).

Ковариационная матрица $f(X)$:

$$Var(f(X)) = \mathbb{E}\Big[\big(f(X) - \mathbb{E}f(X)\big)\big(f(X) - \mathbb{E}f(X)\big)^\top\Big].$$

Справедливо матричное неравенство

$$Var(f(X)) \preceq \mathbb{E}\Big[J_f(X)\,\Sigma\,J_f(X)^\top\Big],$$

где $\preceq$ означает порядок Лёвнера: $A \preceq B \Leftrightarrow B - A$ положительно полуопределена.

Выделение носителя в пространствах высокой размерности

$$z_i = f(x_i) \in \mathbb{R}^d$$

$$m_j = \min_i z_{ij}, \quad M_j = \max_i z_{ij}, \quad \Delta_j = M_j - m_j$$

$$V_K = \prod_{j:\,\Delta_j > 0} (\Delta_j)$$

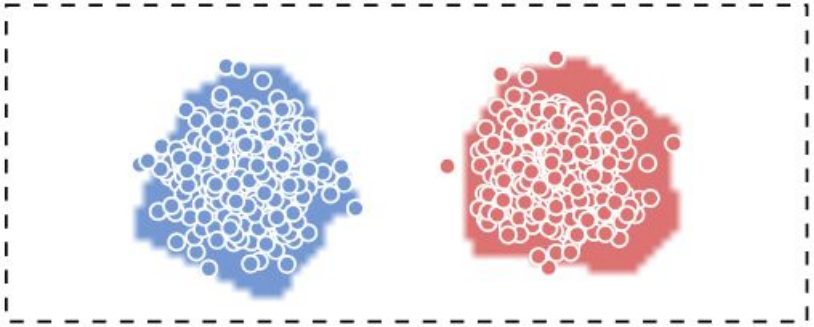
$$\Sigma_i = J_f(x_i)\,\Sigma\,J_f(x_i)^T$$

$$S = \sum_{i=1}^{N} \sqrt{\det(\Sigma_i)}$$

$$\rho_{avg-opt} = \sum_{c=1}^{C} \frac{N_c}{N} \cdot \frac{N_c}{\sum_{i \in c} \sqrt{\det(\Sigma_i)}}$$

$$N_{\text{noise}} = \rho_{avg} \cdot V_K$$

$$N_{\text{noise-opt}} = \sum_{c=1}^{C} \frac{N_c}{N} \cdot \frac{N_c}{\sum_{i \in c} \sqrt{\det\left(J_f(x_i)\,\Sigma\,J_f(x_i)^T\right)}} \cdot \prod_{j:\,\Delta_j > 0} (\Delta_j)$$



Выделение носителя в пространствах высокой размерности

Таблица 14 — Оценка снижение эффективности атаки на наборе данных CIFAR-10, модель ResNet-18 и ResNet-56. Все метрики представлены в процентах (%)

Модель	Точность (↑)	ASR-PGD (↓)
ResNet-18	80.6 ± 3.0	14.3 ± 3.5
ResNet-56	85.2 ± 2.5	98 ± 0.4

Таблица 15 — Точность детекции OOD на наборе данных CIFAR-10, модель ResNet-18 и ResNet-56. Все метрики представлены в процентах (%)

Модель	Точность (↑)	OOD-noise (↑)	OOD-class (↑)	OOD-PGD (↑)
ResNet-18	80.6 ± 3.0	100 ± 0	89.8 ± 1.5	83.4 ± 2.2
ResNet-56	85.2 ± 2.5	14.2 ± 4.3	2.6 ± 0.8	0 ± 0

Метод обучения для предотвращения избыточного сжатия

$$\mathcal{L}_{new}(\theta) = \mathcal{L}_{CE}(\theta) + \lambda_1 \mathcal{L}_{marg}(\theta) + \lambda_2 \mathcal{L}_J(\theta)$$

$$\mathcal{L}_{CE}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y_i \mid x_i)$$

$$a(i) = \frac{1}{|C_{y_i}| - 1} \sum_{\substack{j \in C_{y_i} \\ j \neq i}} \|z_i - z_j\|$$

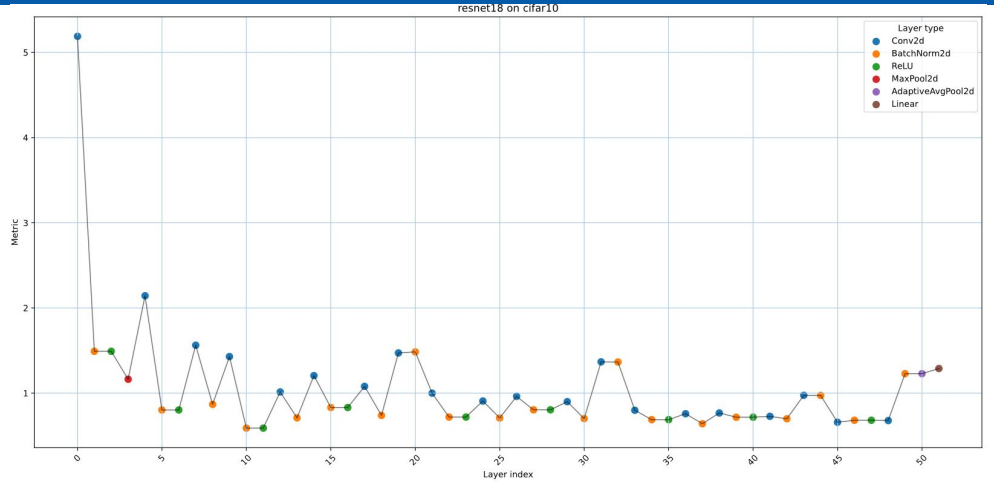
$$b(i) = \min_{k \neq y_i} \frac{1}{|C_k|} \sum_{j \in C_k} \|z_i - z_j\|$$

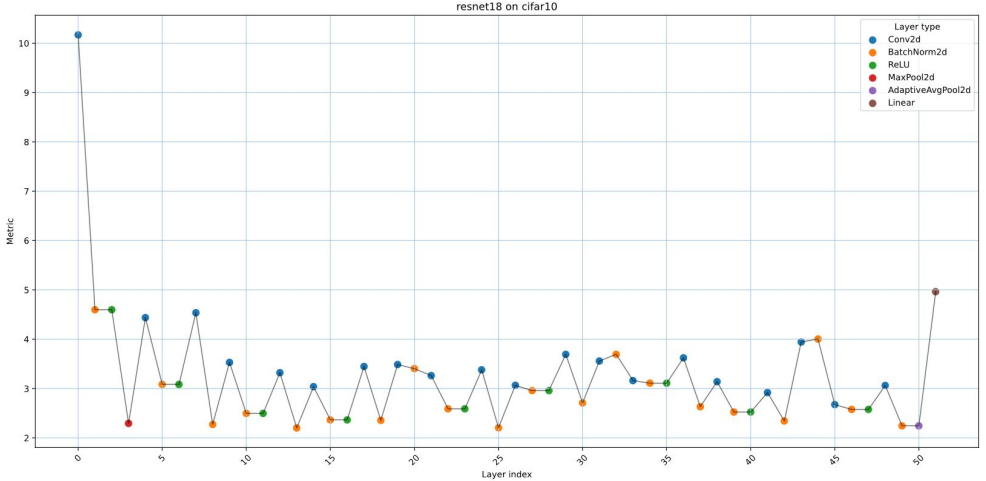
$$s(i) = (max(0, a(i) - b(i) + \delta)$$

$$\mathcal{L}_{marg}(\theta) = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

$$J_f(x) = \frac{\partial f_\theta(x)}{\partial x}$$

$$\mathcal{L}_J(\theta) = \frac{1}{C} \sum_{c=1}^{C} \left(\frac{1}{N_c} \sum_{i \in c} \mathrm{Tr}\left(J_f(x_i) \, \Sigma \, J_f(x_i)^T \right) - \tau \right)^2$$





Выделение носителя в пространствах высокой размерности

Таблица 16 — Оценка снижение эффективности атаки на наборе данных CIFAR-10, модель ResNet-56 с двумя разными функциями потерь. Все метрики представлены в процентах (%)

Функция потерь	Точность (↑)	ASR-PGD (↓)	ASR-MMA (↓)
$\mathcal{L}_{CE}$	85.2 ± 2.5	98 ± 0.4	89 ± 1.4
$\mathcal{L}_{new}$	81.7 ± 4.3	17.7 ± 5.1	4.2 ± 2.9

Таблица 17 — Точность детекции OOD на наборе данных CIFAR-10, модель ResNet-56 с двумя разными функциями потерь. Все метрики представлены в процентах (%)

Функция потерь	Точность (↑)	OOD-noise (↑)	OOD-class (↑)	OOD-PGD (↑)
$\mathcal{L}_{CE}$	85.2 ± 2.5	14.2 ± 4.3	2.6 ± 0.8	0 ± 0
$\mathcal{L}_{new}$	81.7 ± 4.3	95.8 ± 1.5	84.6 ± 3	75.6 ± 7.2